Module 3: Assignment 1

Yanyang Pei (B3)

Introduction

Hypothesis: chemical parameters malic and CO2 can classify white wine while lactic acid and volatile can classify red wine. NMR data is associated with chemical measurements, but cannot fully represent it.

Materials and Methods

Data analyses were done with Python 3.1.2. Two data measured for the same samples were used, where one contains the peak intensity of NMR analysis and the other contains chemical measurement. Grubbs test was used to help preprocess the data, and empty was removed or filled in.

Studies of the potential relationship between two datasets were performed by using PLS followed by modeling. The result was compared with PCA analysis and decision boundary plot.

The model was generated by PLS regression as well as logistic regression, and classification ability was studied by plotting confusion matrix and decision boundary plots.

Results

Preprocess and prestudy

Grubbs' test identified three outliers in the chemical measurement data. Python detected missing values, with two in the NMR data and one in the chemical data. Rows with missing NMR data, along with their corresponding rows in the chemical data, were removed. The missing value in the chemical data, as well as its outlier, were replaced by the average. The modifications are summarized in Table 1.

PLS was used to explore the correlation between the two datasets. The NMR data was used as the explanatory variable, while the chemical measurements were used as the response variable. The plot is shown in Figure 1, and a plot using dummy variables was shown in Figure A.1 in Appendix.

Table 1. Pre-processing of data.

ID	Column	Problem	Replacement
14	Lactic acid	Outlier: 10.077	1.0323
37	NMR210	Missing data	Removed whole row
27	NMR269	Missing data	Removed whole row
24	Density	Missing data	0.9951
19	Ethyl Acetate	Outlier: 1543	486.2850

12	CO2	Outlier: 1363	803.6825

PLS regression and model study

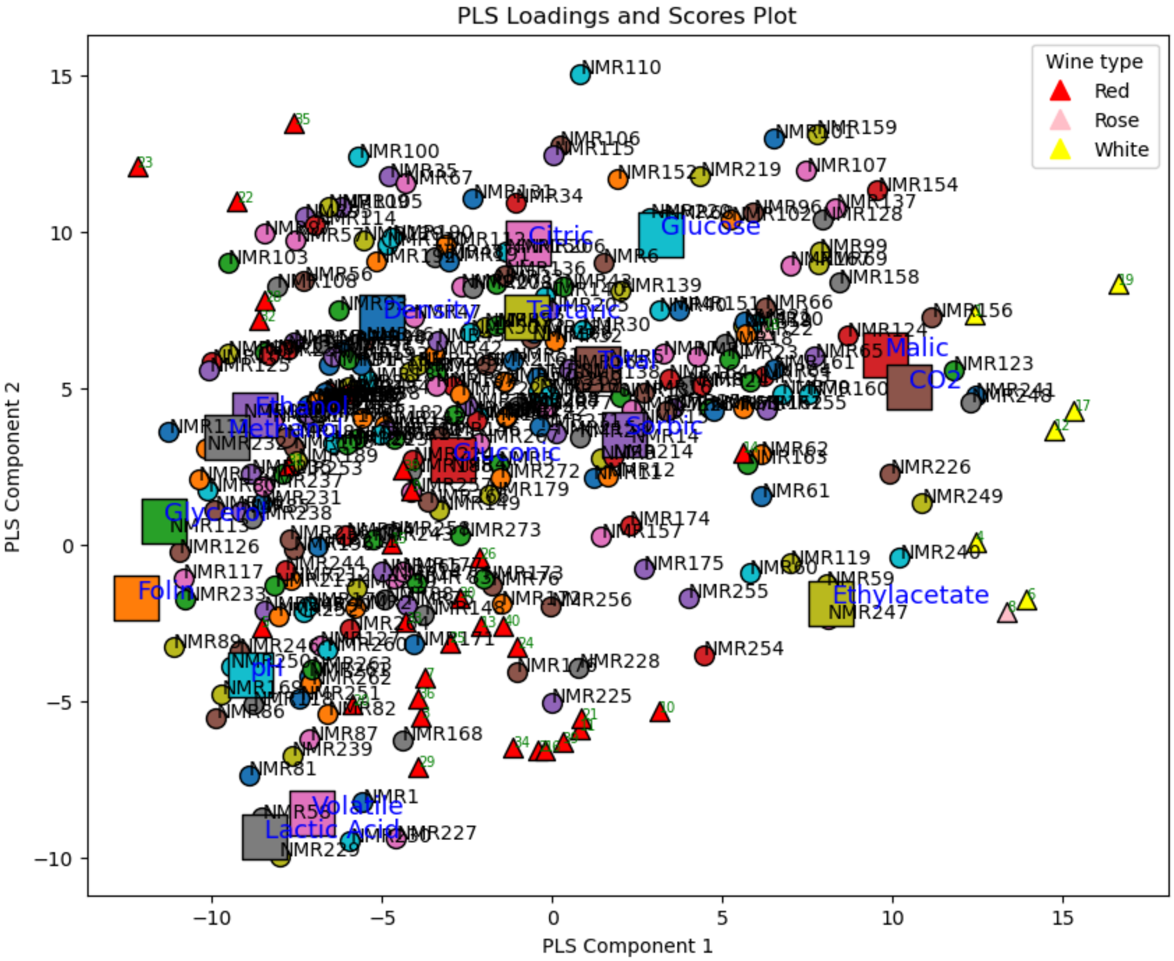
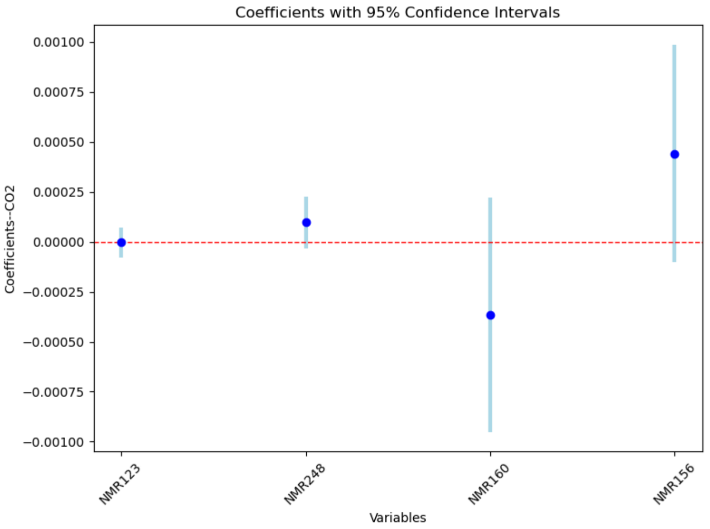


Figure 1. PLS plot illustrating the relationship between datasets: triangles represent different wine types, circles represent NMR data points, and squares correspond to chemical measurement parameters.

It can be seen that white wine is closely clustered with property malic and CO2, and slightly grouped with ethyl acetate, while red wine spreads widely out and is closely related to volatile and lactic acid, and gluconic. However, wine rose has low research value since there’s only two points and there’s large variation between it. The five parameters and their potential correlated

NMR signal are studied by making effect plots, shown in Figure 2 - Figure 5, and the results are summarized in Table 2.



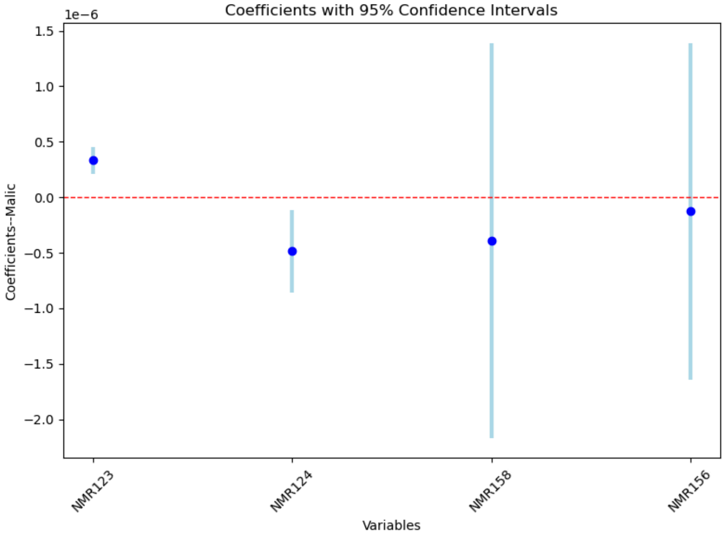
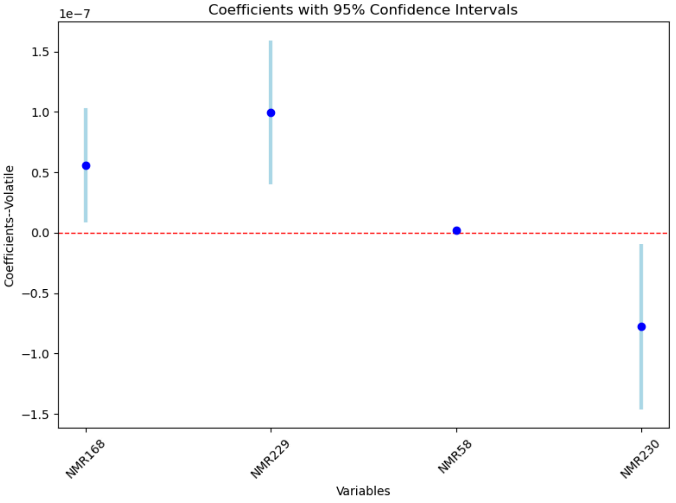


Figure 2. Effect plot for CO2.

Figure 3. Effect plot for malic.



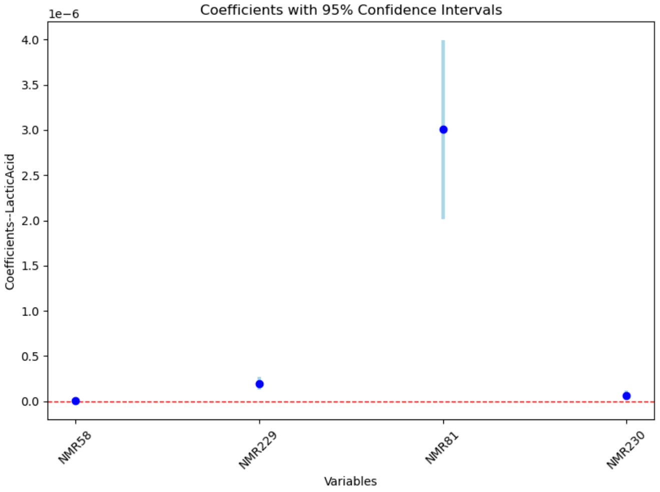


Figure 4. Effect plot of volatile.

Figure 5. Effect plot of lactic acid.

Table 2. Correlation and significance of chemical parameters and NMR signals, where blue color indicates negative correlation and red indicates positive correlation.

Wine type	Chemical parameter	Correlated NMR signal	Correlation and Significance
White	CO2	NMR123	ns
		NMR248	ns
		NMR160	ns
		NMR156	ns
	Malic	NMR124	*
		NMR158	ns
		NMR123	*
		NMR156	ns
Red	Volatile	NMR168	ns
		NMR229	ns
		NMR230	ns
		NMR58	ns
	Lactic acid	NMR229	*
		NMR58	ns
		NMR230	*
		NMR81	*

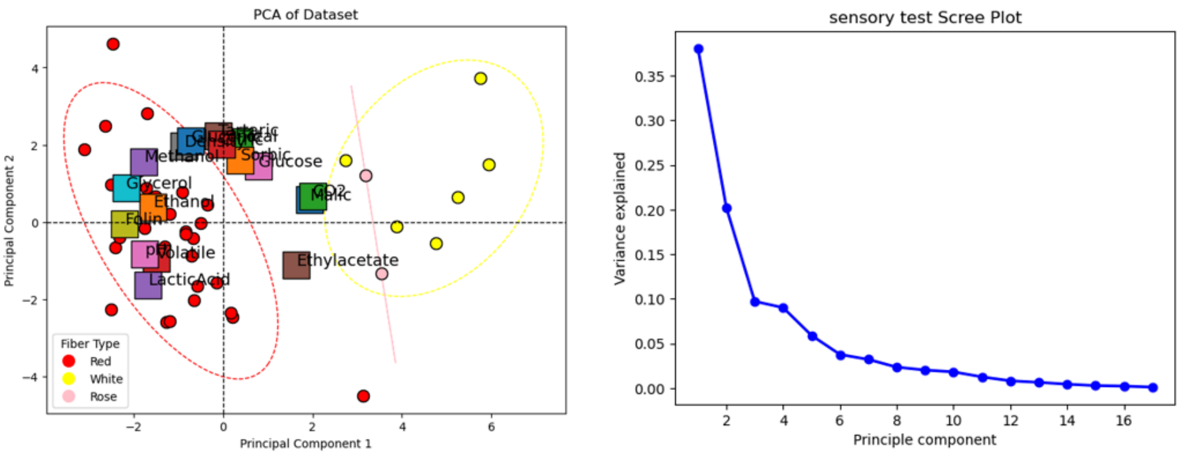


Figure 6. PCA biplot and scree plot of chemical measurements.

The PCA analysis of the chemical data aligns well with the relationships observed in the previous PLS study. Red wine is strongly associated with pH and volatility, while white wine shows a correlation with CO2 and malic acid. However, the scree plot reveals that PC1, PC2, and PC3 explain relatively low amounts of variance, suggesting that the system is quite complex.

Besides, malic and lactic acid were chosen to represent the other parameters (CO2 and volatile) to do the decision boundary shown in Figure 7. From which it can be concluded that malic is associated with white wine and lactic acid is well associated with red wine. This aligns well with the PLS study.

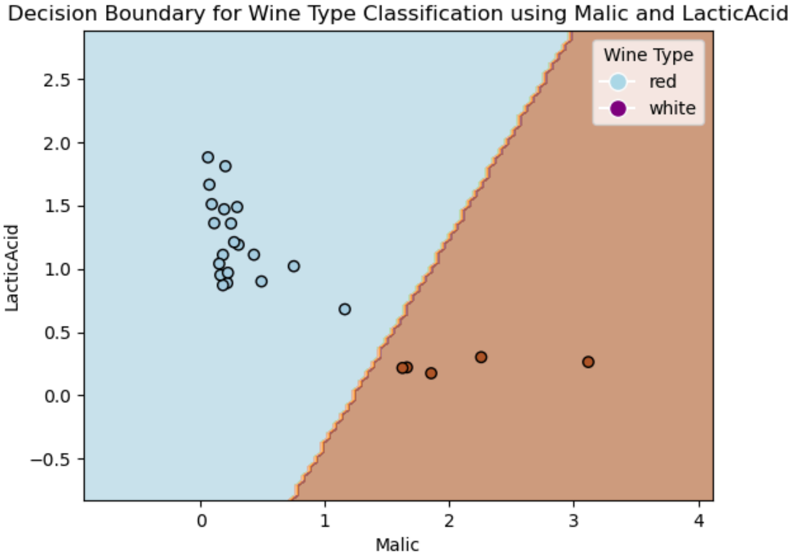


Figure 7. Decision boundary for determining association with parameters and wine type.

Classification of model

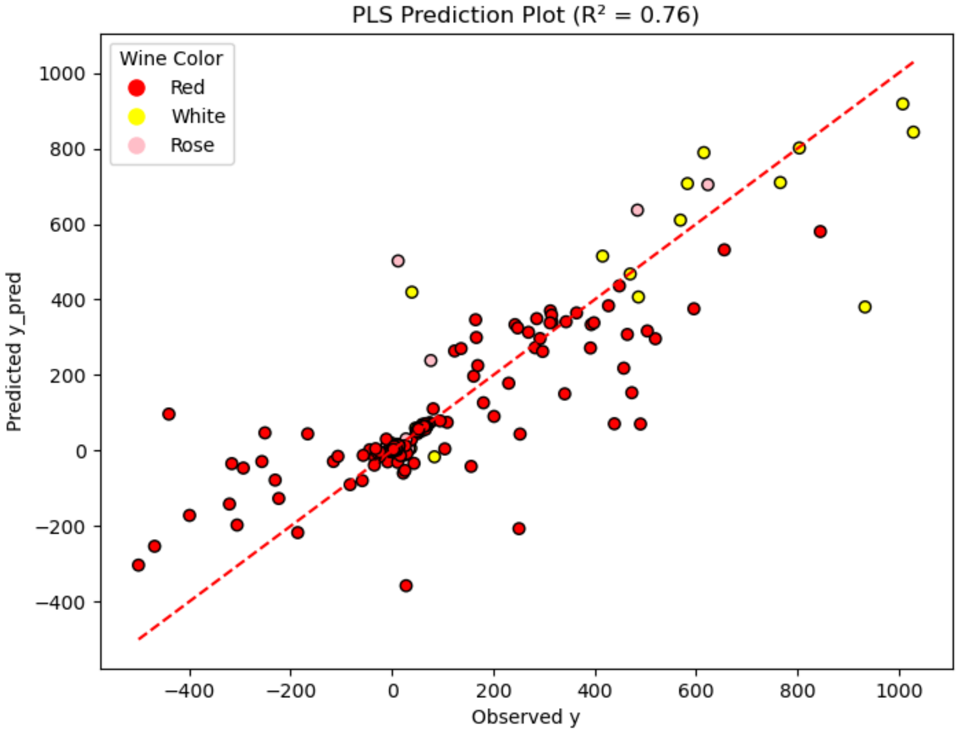


Figure 8. PLSRegression, where R2 is 0.76 and points are colored by wine type.

The PLS regression plot shown in Figure 8 shows a R² value of 0.76, suggesting a good fit of model to data. This indicates a strong relationship between NMR-predicted and observed chemical measurements, demonstrating that NMR data can effectively explain the chemical features. Red wines (red dots) appear to be predicted more accurately compared to white wines (yellow dots) and rose wines (light pink dots). The white and rose wines exhibit more variability, indicating that the model may have more difficulty predicting these wine types.

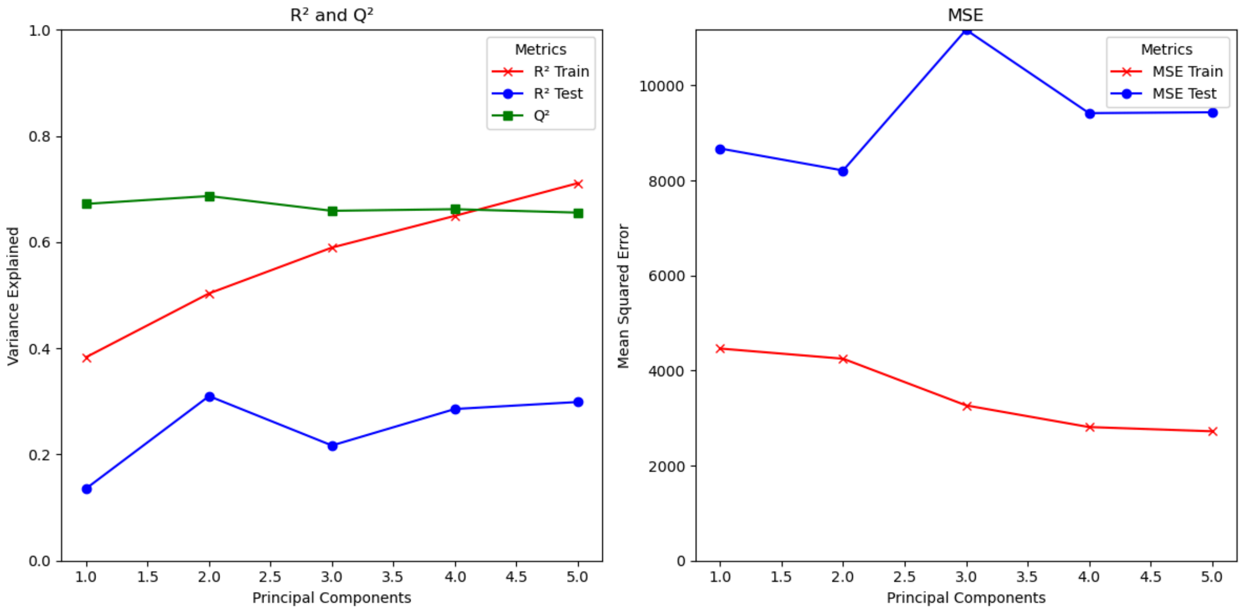


Figure 9. R^2, Q^2 and MSE plot with changing principal components number

The plot shown in Figure 9 illustrates that after the second principal component, the R² for the test data begins to decrease while the MSE increases, indicating model overfitting. This trend aligns with the change of Q² value, which also demonstrates a decline in the model's performance after two principal components.

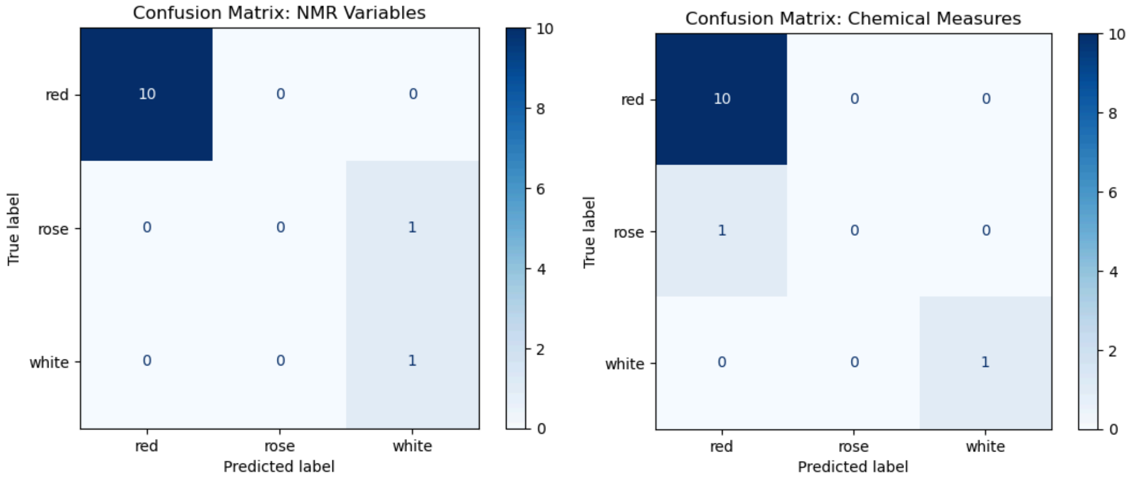


Figure 10. Confusion matrix for NMR data (left) and chemical measurements data (right)

Confusion plots shown in Figure 10 were made to study the classification property of the data. From which it can be seen that, 10 out of the 10 red wine samples and one white wine sample for both NMR and chemical measure are correctly classified, while all two rose samples are misclassified. This indicates that the model using NMR variables performed well in classifying red wines but struggled with classifying roses. What's more, even though the model classifies correctly for white wine in this test, the small amount of data causes challenges in categorizing it.

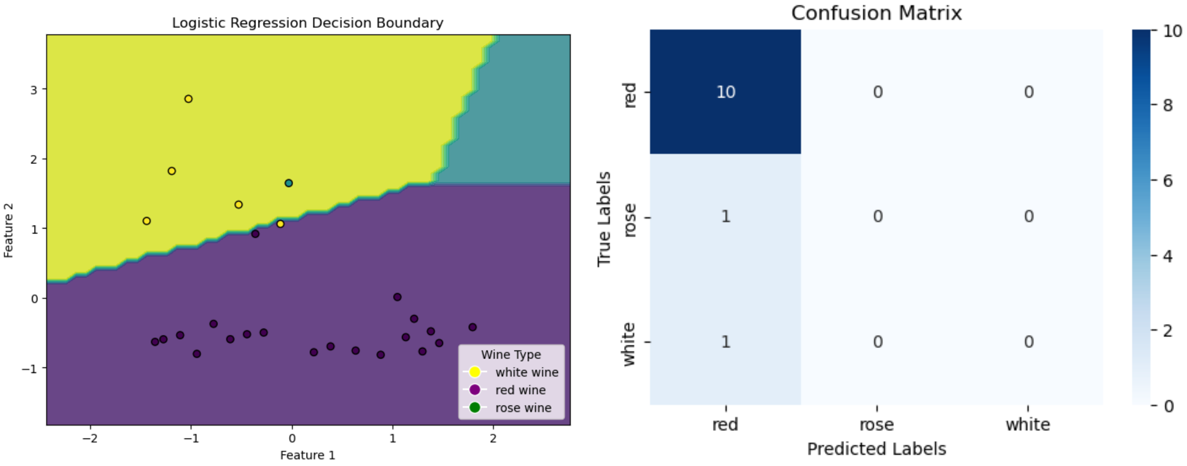


Figure 11. Logistic regression decision boundary classifies with wine type and confusion matrix for the model of chemical measurements.

By using logistic regression, a cross-validation score of 0.9333 and accuracy of 0.9167 was found. This suggests that the model generalizes well for both training and testing data. The confusion matrix shown in Figure 11 indicates that only 1 out of 11 red wine samples was misclassified, but both white wine and rose wine have very little sample number and are not well classified. The classification report shown in Figure A.2 in Appendix also reveals the same result, and besides that, the macro average shows a mixed performance across the classes due to the misclassification of rose wine.

The decision boundary plot was divided into three regions, where purple represents red wine prediction and yellow represents white wine prediction. A good separation between red wine and white wine can be seen. However, rose wine is close to the decision boundary.

The logistic regression and model study for NMR data shows a very similar cross-validation score and accuracy. Same results are driven from the decision boundary plot as well as the confusion matrix, this is shown in Figure A.3 in Appendix.

Discussion

Although NMR data and chemical measurements show close correlation in PLS study, it appears that they cannot fully represent each other.

The result from PLS is well aligned with the effect plot and PCA study. Also the decision boundary plot shows the same information that those chemical parameters chosen (malic, CO2, volatile and lactic acid) can well classify the wine types.

The model generated can accurately classify results for red wine. But for white wine and rose, due to not enough data, the model fails to give a precise classification.

Future study can be done on studying more relationships between chemical measurements and NMR data. Evenmore, studying the similarity between rose wine and red wine.

Conclusion

In summary, both chemical measurements and NMR data performed similarly in classifying wine types, with slight differences in generalizability. The decision boundary plots show that Malic and Lactic Acid provide a reasonable separation between red and white wines, but additional features and a more balanced dataset would be required to improve the classification of rose wines. The results underline the importance of accurate feature selection and data balancing when applying classification models to wine types.

Appendix

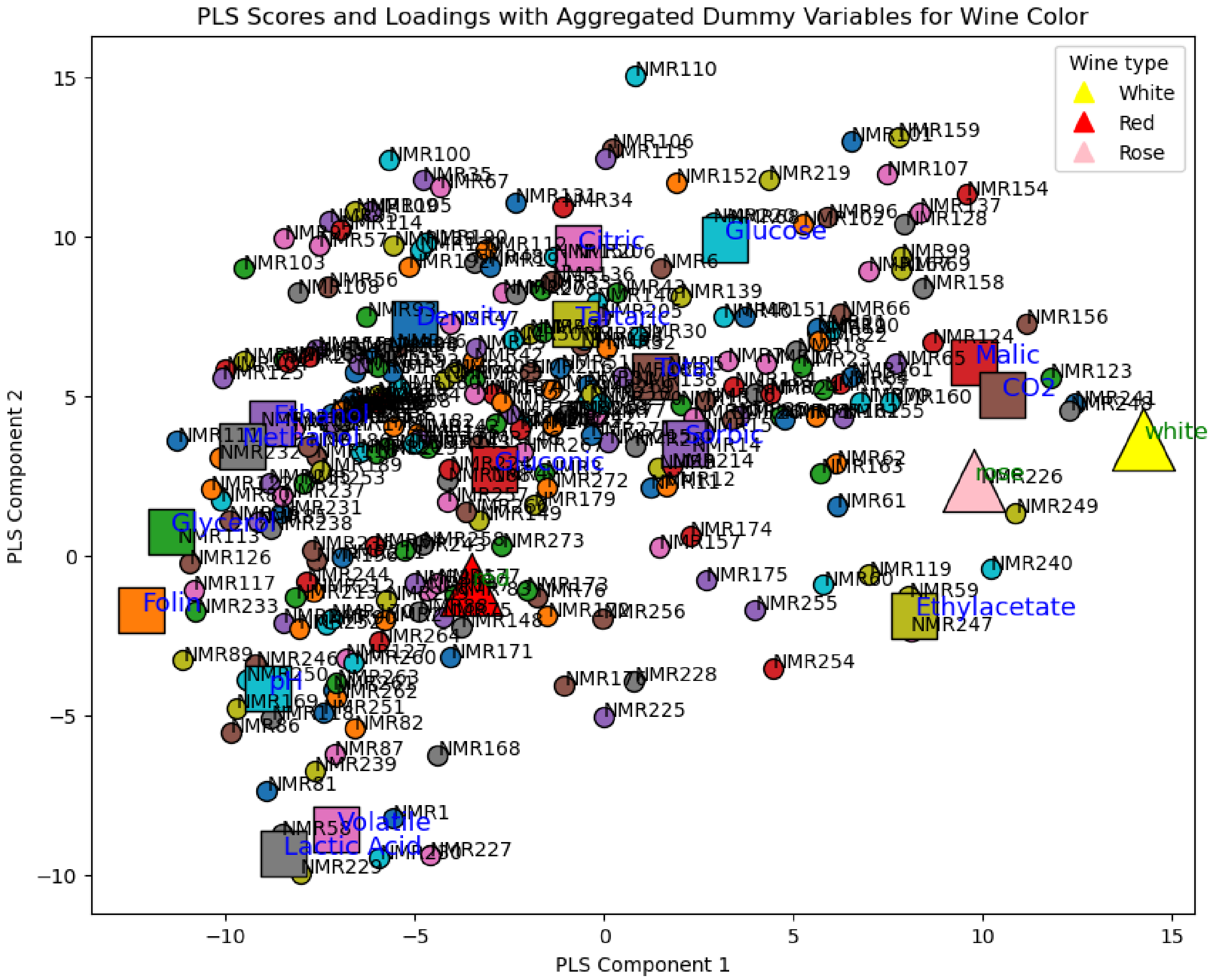
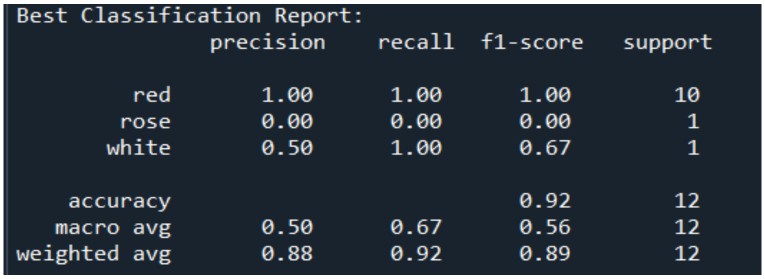


Figure A.1 PLS plot using dummy variables


```
Best Classification Report:
              precision    recall  f1-score   support

         red       1.00      1.00      1.00        10
        rose       0.00      0.00      0.00         1
       white       0.50      1.00      0.67         1

    accuracy                           0.92        12
   macro avg       0.50      0.67      0.56        12
weighted avg       0.88      0.92      0.89        12
```

Figure A.2. Classification report of chemical measurements data.

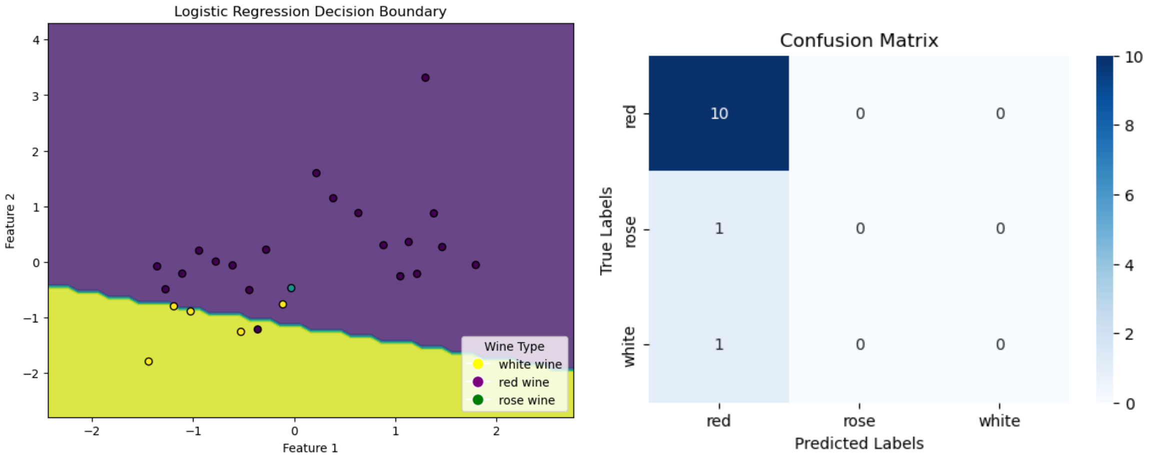


Figure A.3. Logistic regression decision boundary classifies with wine type and confusion matrix for the model of NMR measurements.

Acknowledgements

AI was used in code writing.